Adaptive Comparative Judgement: A Tool to Support Students' Assessment Literacy

Adaptive Comparative Judgement: A Tool to Support Students' Assessment Literacy

Susan M. Rhind. Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVMS, PhD, FRCPath, PFHEA, MRCVS, Director of Veterinary Teaching, Chair of Veterinary Medical Education and Assistant Principal (Assessment and Feedback). Specific interests in assessment and feedback, e-learning, curriculum development and student well-being/ support. E-mail: susan.rhind@ed.ac.uk.

Kirsty J. Hughes. Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVM&S, BSc, MSc, PhD, FHEA, MRCVS, Research Assistant in Veterinary Medical Education. Specific interests include assessment and feedback, e-learning, the student experience and staff development.

Donald Yool. Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVMS, PhD, DipECVS, CertSAS, MRCVS is Head of Companion Animal Soft Tissue Surgery Specific interests in assessment and in use of digital media in supporting undergraduate surgical teaching.

Darren Shaw. Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BSc, PhD, Senior Lecturer in Epidemiology. Specific interests in quantitative epidemiology of zoonotic / purely veterinary diseases as well as the clinical epidemiology associated with animal health that is of veterinary importance.

Wesley Kerr. Information Services at the University of Edinburgh, 19 Buccleuch Place, Edinburgh EH8 9LN. GRSC, BSc, PGCert. Senior eLearning advisor with the Educational Design and Engagement team. Specific interests include assessment, feedback and student engagement.

Nicki Reed. Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVM&S, Cert VR, DSAM (Feline), Diploma ECVIM-CA, MRCVS. Senior Lecturer in Internal Medicine. Specific interest in teaching clinical application of theoretical knowledge. *NR's current address is Lumbry Park Veterinary Specialists, Lumbry Park, Selborne Road, Alton, Hampshire GU34 3HF.

Abstract

Comparative judgement in assessment is a process whereby repeated comparison of two items (e.g. assessment answers) can allow an accurate ranking of all the submissions to be achieved. In Adaptive Comparative Judgement (ACJ), technology is used to automate the process and present pairs of pieces of work over iterative cycles. An on-line ACJ system was used to present students with work prepared by a previous cohort at the same stage of their studies. Objective marks given to the work by experienced faculty were compared to the rankings given to the work by a cohort of veterinary students (n=154). Each student was required to review and judge 20 answers provided by the previous cohort to a free text short answer question. The time that students spent on the judgement tasks was recorded and students were asked to reflect on their experiences after engaging with the task. There was a strong positive correlation between student ranking and faculty marking. A weak positive correlation was found between the time students spent on the judgements and their performance on the part of their own examination which contained questions in the same format. Slightly less than half of the students agreed that the exercise was a good use of their time, but 78% agreed that they had learnt from the process. Qualitative data highlighted different levels of benefit from the simplest aspect of learning more about the topic to an appreciation of the more generic lessons to be learned.

Key Words: assessment literacy, peer assessment, feedback

Background

The concept of comparative judgement in assessment is not new, having first been described by Thurstone in 1927 [1] as a process whereby repeated comparison of two items, if carried out a sufficient number of times across all the items, can allow an accurate ranking of these items to be achieved. In 2012, Pollit, extended this concept to one of 'Adaptive Comparative Judgement' (ACJ) by describing a system where technology can be used to automate the operation of the underlying algorithm and present judges with pairs of pieces of work for comparison over iterative cycles [2]. Through these iterative cycles of judgement rather than assigning of marks to pieces of work, all the items are ultimately sorted into a rank order [3]. It is then possible for assessors to agree an aligned marking scheme if required or simply provide students with the rank ordering as feedback on their performance. Whilst

ACJ has mainly been described in the context of faculty judging students work, there is also the potential to explore the principle in the context of peer assessment.

It has been shown that peer assessment, if well supported can help students develop their own ability to self-assess [4]. By giving students the opportunity to act as judges, there are potential additional benefits to them in terms of increasing their skills in understanding the range of quality in other students work. Smith et al (2013) presented a study with business students showing that the development of students' ability to judge standards of performance on student work correlated with enhanced marks [5]. This type of activity is being used in other disciplines in Higher Education to build students 'assessment literacy' skills, a term encompassing the range of knowledge, skills and attributes necessary to understand the purpose and process of assessment [6, 7].

Developing skills in self-assessment requires practice, time and the opportunity to consider the strengths and weaknesses of a range of pieces of assessed work. This aspect has been emphasized by Sadler who in particular has championed the notion of provision of 'substantial evaluative experience' as a core part of the design of a curriculum [8]. Boud has also published frequently in this area, emphasizing the importance of students becoming assessors in order to fully understand the nature of good quality work [9-11].

In the study reported here, we sought to compare the objective marks (and implicitly the ranking) given to pieces of work by experienced faculty to the rankings given to the work by a larger group of veterinary students who had recently learned the material relevant to the question. The students were engaging in the ranking process as part of a curriculum intervention to build skills in assessment literacy and allow them an explicit opportunity to experience a range of pieces of work of different quality from a previous cohort at the same stage in the curriculum.

In carrying out this study we sought to explore the following hypotheses:

1. Students' abilities as markers will positively correlate with actual faculty marks given to pieces of work.
2. Students' will find engaging with a range of pieces of work of differing standards helpful beyond the context of the specific course in the study.

Methods

Context

The context of this study was a small animal medicine course in the second or third year of a veterinary degree programme. Students in the cohort under study included students with previous degrees on an accelerated 4 year programme and students on a five year programme entering straight from high school. The cohort size was 154. The course is examined by in course assessment) and an end of course degree examination comprising MCQs and short answer questions.

Ranking activity

The answers to an examination question from the previous cohort (n=162) were transcribed and entered into a commercially available system (Digital Assess[a]) which allows pairs of pieces of work to be presented to assessors who then compare the work and select which of the 2 is better using the principle of Adaptive Comparative Judgement. As the students work through their comparisons, the system fine tunes the comparisons to focus on those that are most similar and ultimately presents a complete rank order of all pieces of work with an associated reliability statistic. The system also collects data on the amount of time spent on both each individual judgement and also overall time spent per student on the judging process.

Students were given an introduction to the system and an explanation of the question they were assessing and an outline expected answer for the question in a plenary session at the start of a lecture. The question had previously been marked by a single faculty member who taught the material and given a mark out of 10. The question selected for the exercise was chosen due to the spread of marks given (consistent with a range of answer quality). The overall mark profile is shown in Figure 1 (Mean mark 4.7, median 5). The question used is shown in Box 1.

[Insert Figure 1 here]

You have made a clinical diagnosis of allergic skin disease in a three year old, male West Highland white terrier called Angus. You have decided to start Angus on a food trial to determine whether a protein in his diet is contributing to his pruritus and inflammation. Angus is 6/10 pruritic but your cytology samples did not reveal any evidence of bacterial or Malassezia skin infections.

A) **LIST** the different options for an appropriate diet in this case.
(3 marks)

B) Briefly discuss how to select an appropriate food for Angus.
(3 marks)

C) Briefly discuss how the diet trial should be conducted to maximise compliance and a diagnostic outcome.
(4 marks)

Box 1

Students were given a 2 week period to perform 10 parallel judgements using the on-line system i.e. each student made judgements on a total of 20 pieces of work. Throughout the rounds of comparisons the system calculates a reliability statistic (between 0 and 1) which was recorded.

Correlation of Faculty and Student Judgements

The final output from the ACJ system ranks all pieces of work in order based on the judgements made. To compare faculty and student opinions on the work, a correlation of the actual marks given by faculty was carried out against the rank order using Spearman Rank correlation.

Performance on the Degree Examination

The final degree examination marks were correlated against the time students spent comparing questions and completing their judgement.

Survey

Following the ranking exercise, students were asked to complete a short on-line questionnaire on their experiences comprising a mixture of Likert scale and open ended questions. Responses to open ended questions were grouped into themes.

Ethical approval for this study was sought and received from the College of Medicine and Veterinary Medicine's Committee for the use of student volunteers at the University of Edinburgh (Ref: 2014/24)

Results

Reliability

The reliability statistic at the end of the 20 rounds of judgement was 0.98 indicating a high level of reliability.

Time spent on judgement

The average time taken per judgement was 139 Seconds; standard deviation 83. Median 126 seconds (Figure 2).

[Insert Figure 2 here]

Student judging compared to faculty judging

There was a positive correlation between faculty mark and ranking by students (0.690, $p<0.001$, Figure 3)

[Insert Figure 3 here]

Performance on Degree Examination with Time spent on Ranking Task

The degree examination was divided into multiple choice elements and short answer and then correlated to the total and average time spent making judgements (Table 1).

[Insert table 1 here}

Student Evaluation

The survey was completed by 67 of the total 154 students (44% response rate). 68% of students were extremely positive or positive about their experience with the software, 24% quite positive and 8% not at all positive.

Responses to a series of Likert scale questions on the intervention are shown in Table 2.

[Insert Table 2 here]

Although the majority (>75%) agreed or strongly agreed that they had learned from what their peers had submitted, slightly less than 50% agreed that it was a good use of their time. A similar percentage (49%) indicated that they would recommend the tool for use in other courses.

Students were asked whether reviewing other people's answers had made them think differently about how they answered questions. 69% answered yes, 31% answered no. Students were then asked to explain their answer in more detail. Themes in this free text could be broadly grouped into those relating to the benefit being around revision of the particular topic and more generic benefits.

Learning more about the specific topic:

> 'Learned about the topic being addressed and also the type of question that can be asked and the answer expected from us.' S4

> 'I learned about the topic while doing the exercise and it was really interesting to see other peoples answers.' S10

Many more comments related to the broader learning around assessment and approach in general; particularly where it highlighted good practice

‘It was interesting to see different ways that students laid out their work. It was also interesting to see different angles from which students approached the questions.’ S3

‘Some people approach things in a very different way to me - that was interesting’ S25

And this point was extended by several students to indicate that reflection on this had helped them with their future approach to similar style of questions:

‘I am typically someone who waffles during exams. While reading the various answers, I was able to pick up useful tips on keep the answers short, simple and to the point.’ S7

‘seeing some of the answers was definitely a bit of "What not to do!" S2

‘It reminded me of the need to use appropriate terms/language during writing answers and that concise, well-informed answers are better than long rambles. S18

Linked to this were several comments on the benefit of being given the examiner perspective

‘This has certainly given me a level of sympathy for graders’ S16

‘Now I know which methods are easier for the professors to grade.’ S20

The main negative issue raised by students in the data was the focus on one question only with students expressing a desire to have more examples of different subject areas to review.

‘It would have been much better if every question was a different question instead of the same one 10 times as this just became boring. And we would have learnt more if different questions were used.’ S2

Limitations

The exercise was carried out in the context of one course in one school and only utilised one question. However the short answer question format is commonly used in many contexts

and schools so results are likely to be of relevance to other educators interested in assessment of free text format. The system of evaluation used was different for faculty and students however it was impractical to have the answers ranked for a second time by the same individual using the ACJ system. The response rate to the survey of 44% is a further limitation.

Discussion

The aim of this study was to facilitate student exposure to work of different quality and engage them in the process of assessing peers work as part of a series of interventions to help support development of assessment literacy skills and understanding of different standards in assessment. We aimed to explore the hypotheses that students' abilities as markers would positively correlate with actual faculty marks given to pieces of work on a subject they had recently covered in the curriculum and that students would find engaging with a range of pieces of work of differing standards helpful beyond the context of the specific course in the study.

We have previously shown that students abilities as judges when assigning marks to work is variable [7]. Although the correlation between the students' ranking and faculty marks was strong in this study, there were clearly outliers where students had ranked an answer markedly different from faculty. Although not possible to explore in the current study design (because the ranking data generated is a cumulative estimate and is not associated with an individual student), it seems a reasonable hypothesis that the students who were more accurate in their ranking, were more academically able.

Student feedback on using the system was mixed with less than half agreeing it was a good use of their time and recommending the exercise for future classes. Despite this, approximately two thirds of students did agree that they had learnt from the process. It would be interesting to explore this dichotomy in future studies through more detailed qualitative approaches such as focus groups. Qualitative data highlighted different levels of benefit from the simplest aspect of learning more about the topic to those who exhibited more metacognitive reflection on the process appreciating the more generic lessons to be learned from such a process.

Furthermore there was a positive correlation between those who spent longer on the judgements and their performance on the part the examination that the exercise was designed to support i.e. free text short answer questions. Whilst this may be a simple

reflection of the more conscientious, able and engaged students spending longer on the task, if we consider the performance on the MCQ part of the examination a 'control' (in terms of a measure of knowledge not requiring skills in description of diagnostic/therapeutic approach), then this suggests that the benefit was more of a direct relation to time spent improving their abilities in the short answer section of the examination. This is consistent with a study reported by Li and Gao, [12] who showed that students who conducted peer assessment performed significantly better than students who did not.

The main negative issue described by the students in the qualitative data was around the focus on one question only with students expressing a desire to have more examples of different subject areas to review. Whilst this would have allowed them to review different topics, the main aim of this study was to give the students insights into differing approaches to answering questions and differing quality work rather than revision of content per se. However given the student feedback discussed earlier, having a system which allowed review of a larger number of topics may well have added to the perceived value of the exercise from the student perspective.

The benefits of peer assessment have been demonstrated elsewhere but often in the context of students marking or giving feedback on one or two other pieces of work [13]. The advantage of the system reported in this study is the ability to electronically facilitate access to a larger number of answers and therefore a wider range of work of differing quality. This approach is consistent with the model of self-regulated learning described by Nicol and Macralane-Dick[14] which highlights the importance of assessment strategies which facilitate self and peer assessment and help clarify what good performance is.

In conclusion, there is evidence that those students who spent longer on carrying out the judgements benefitted when it came to the short answer section of the assessment. Whilst it is not possible to state definitely that this is causative, the lack of correlation with the other aspect of the assessment (the MCQ) suggests it gave these students an advantage. As the survey was anonymous it is not possible to say whether those who enjoyed and engaged with the process more were the ones who benefitted more.

This study provides further support for the utility of assessment literacy interventions in helping students understand more about the assessment process, quality, standards and the challenges assessors face. In particular, the ability of such interventions to encourage students to reflect on their own assessment practice and learn from others was highlighted as a major benefit.

References

[a] http://digitalassess.com/

1 Thurstone LL. A law of comparative judgment. Psychological Review 34:273-86 1927.
2 Pollitt A. Comparative judgement for assessment. International Journal of Technology and Design Education 22(2):157-70, 2012.
3 Whitehouse C, Pollitt A. Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment. < https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf >. The Assessment and Qualifications Alliance (AQA), 2012.
4 Reinholz D. The assessment cycle: a model for learning through peer assessment. Assessment & Evaluation in Higher Education 41(2):301-15, 2016.
5 Smith CD, Worsfold K, Davies L, Fisher R, McPhail R. Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. Assessment & Evaluation in Higher Education 38(1):44-60, 2013.
6 Price M, Rust C, O'Donovan B, Handley K, Bryant R. *Assessment Literacy: The Foundation for Improving Student Learning*. Oxford: The Oxford Centre for Staff and Learning Development; 2012.
7 Rhind SM, Patterson J. Assessment Literacy: Definition, Implementation, and Implications. Journal of Veterinary Medical Education 42(1):28-35, 2015.
8 Sadler DR. Beyond feedback: developing student capability in complex appraisal. Assessment & Evaluation in Higher Education 35(5):535-50, 2010.
9 Boud D, Falchikov N. Quantitative studies of student self-assessment in higher-education - a critical analysis of findings. Higher Education 18(5):529-49, 1989.
10 Boud D, Falchikov N. Aligning assessment with long term learning. Assessment & Evaluation in Higher Education 31(4):399-413, 2006.
11 Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design. Assessment & Evaluation in Higher Education 38(6):698-712, 2013.
12 Li L, Gai F. The effect of peer assessment on project performance of students at different learning levels, . Assessment & Evaluation in Higher Education, 2015.
13 Mostert M, Snowball JD. Where angels fear to tread: online peer-assessment in a large first-year class. Assessment & Evaluation in Higher Education 38(6):674-86, 2013.
14 Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education 31(2):199-218, 2006.

Figure legends

Figure 1

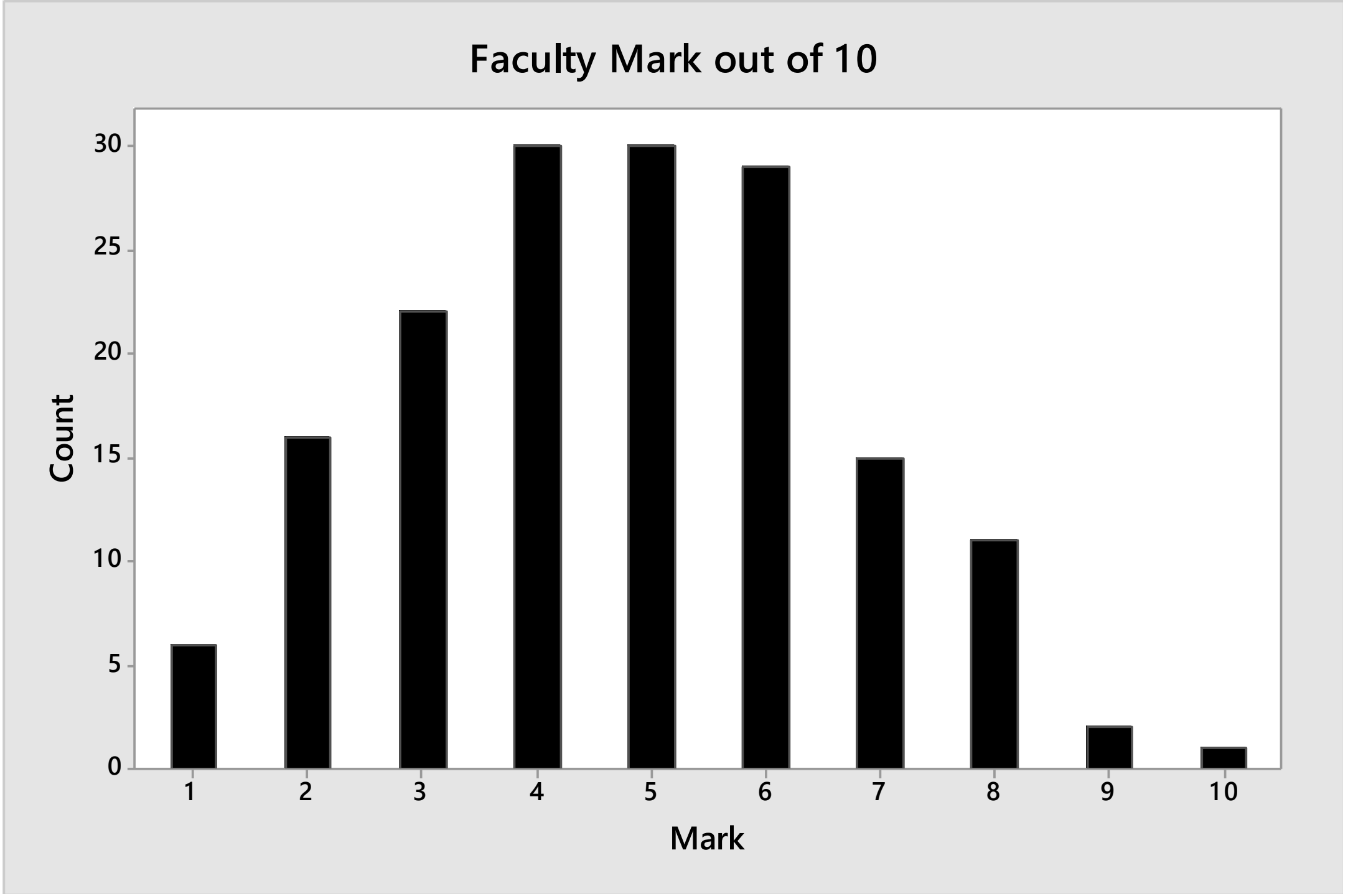
Range of marks given by faculty to the question selected for the adaptive comparative judgement exercise.

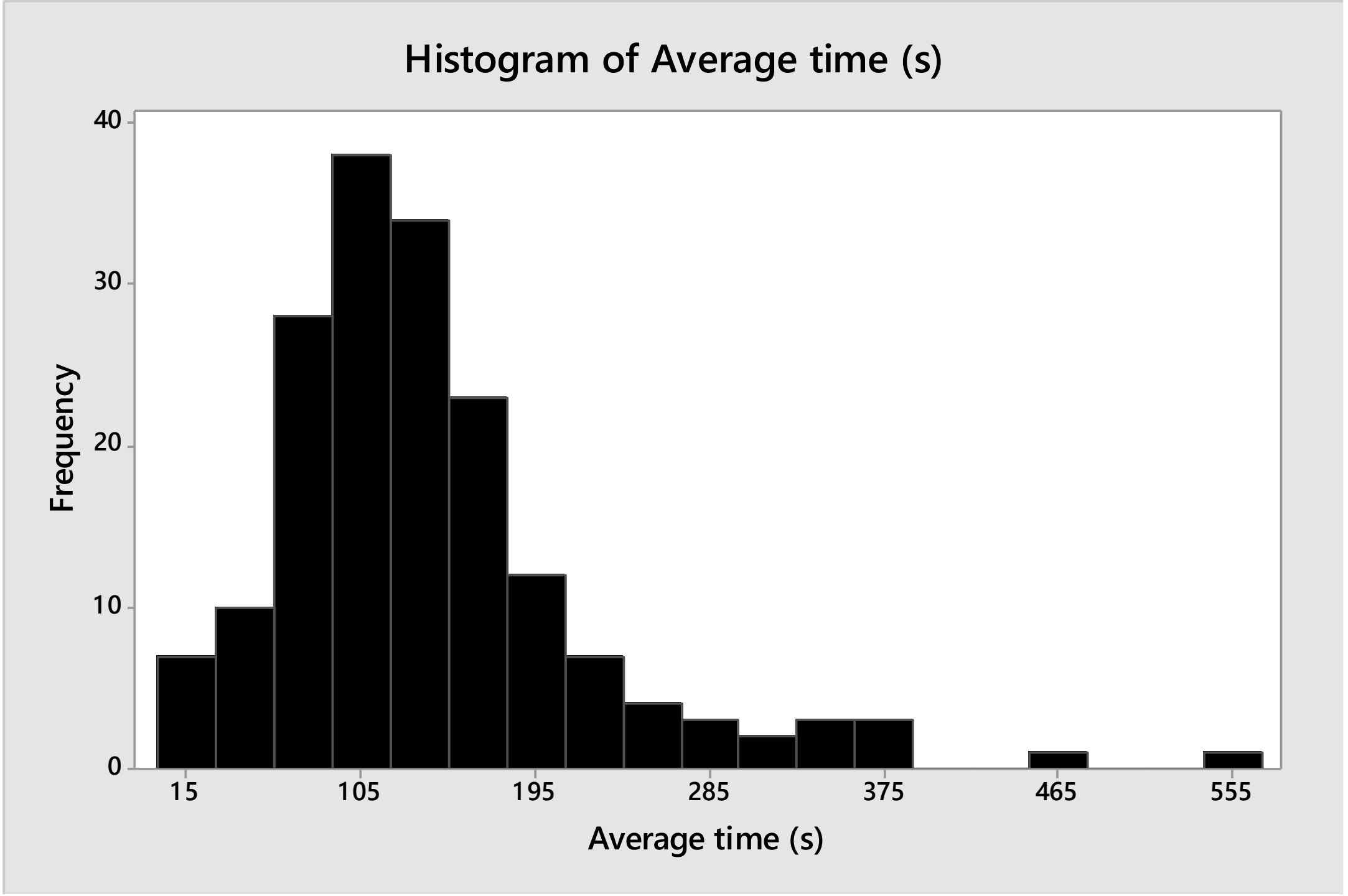
Figure 2

Average time in seconds, spent per student on the judging process captured by the on-line system

Figure 3

Correlation of the ranking produced from the cumulative student judgements with the actual faculty marks given to the pieces of work. Pearson correlation = 0.696, P-Value = 0.000


Faculty Mark out of 10
Count
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
Mark


Histogram of Average time (s)
Frequency
40
30
20
10
0
15
105
195
285
375
465
555
Average time (s)

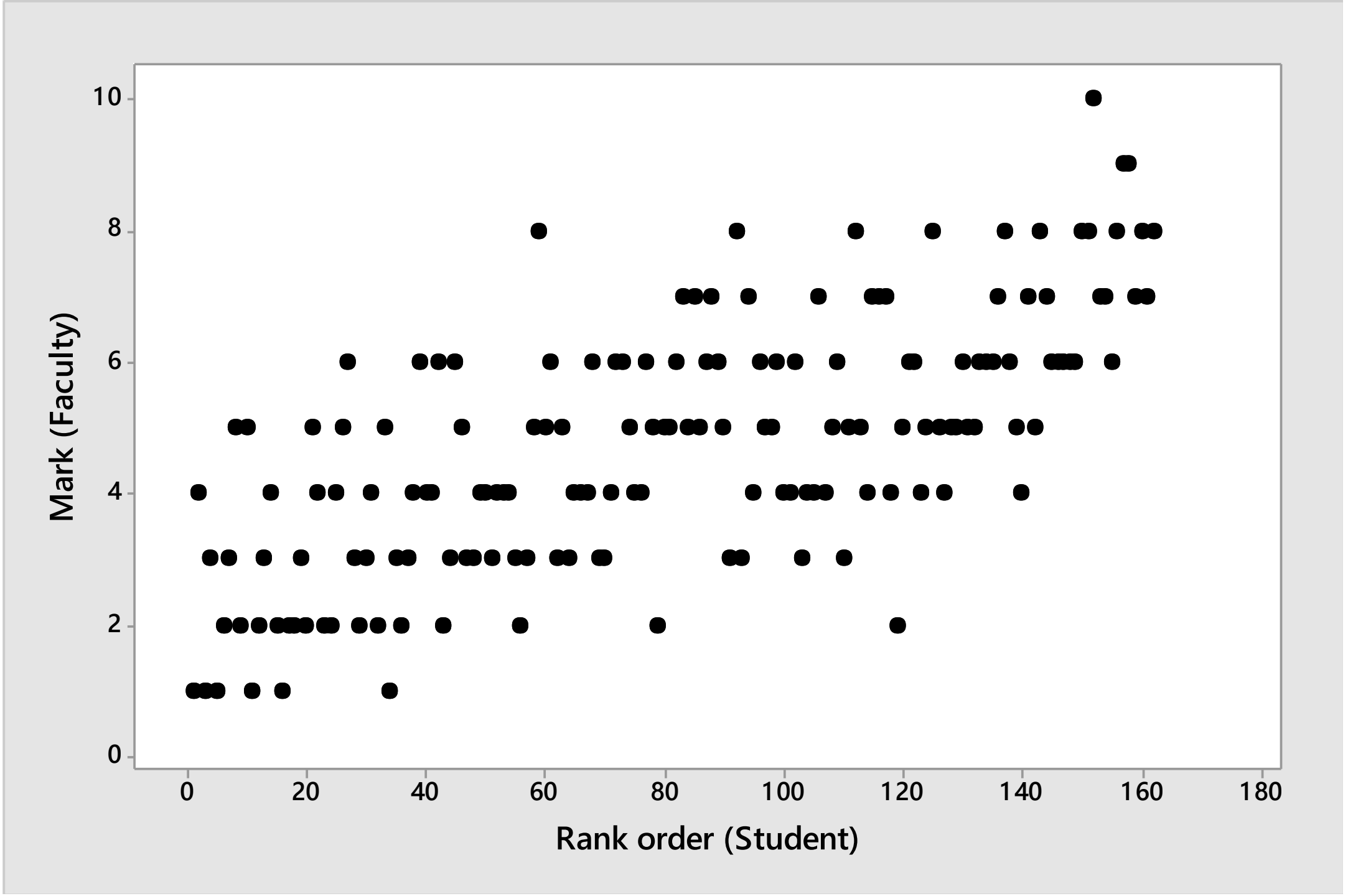

Mark (Faculty)
10
8
6
4
2
0
0
20
40
60
80
100
120
140
160
180
Rank order (Student)

Table 1

Correlation			p-value
Total time spent	MCQ result	0.018	0.153
Average time spent	MCQ result	0.120	0.145
Total time spent	SAQ exam result	0.193*	0.018
Average time spent	SAQ exam result	0.195*	0.017

Spearman rank correlations and associated P values between total and average time students spent making judgements and their performance on the multiple choice (MCQ) and short answer (SAQ) sections of the examination. Significant correlations are asterisked ($p<0.05$)

Table 2

Survey Statement	Strongly Disagree	Disagree	No Strong Feelings	Agree	Strongly Agree
The judging process was enjoyable	4.5	7.5	46.3	37.3	4.5
I have learned from reading what my peers submitted	3	10.4	9	62.7	14.9
This has helped me to understand what markers are looking for	4.5	10.4	16.4	53.7	14.9
This was a good use of my time	4.5	19.4	26.9	49.3	0
I would recommend this for other courses	6	14.9	29.9	43.3	6

Percentage responses to a series of Likert scale questions from the post intervention questionnaire (n=67).